IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10507

Text mining and its Techniques, Applications: An Overview

Jhonathan Quillo-Espino[1], Rosa-María Romero-González[2]

Assistant professor, Faculty of Computer Sciences, Autonomus University of Querétaro, Santiago de Querétaro, Querétaro 76230, Mexico[1]

Professor, Faculty of Computer Sciences, Autonomus University of Querétaro, Santiago de Querétaro, Quéretaro 76230, Mexico[2]

Abstract: Text mining has become a research field that incorporates different techniques such as information retrieval, information extraction, natural language processing, among others, allowing the discovery of important information patterns. This research analyses and describes in a clear and detailed way the elements that comprise it, also shows the most frequent applications that exist today.

Keywords: Text Mining, pre-processes, Text Mining Applications, Text Mining Techniques.

I. INTRODUCTION

The evolution of mankind has made it possible to generate new ways of performing daily tasks in a simpler and more pleasant way. The excessive growth of telecommunications, networks, banks, information, internet, news, technology companies, has led to the production of large amounts of information every day. Every virtual digital life for every person, every organization, every school, etc. A digital environment is appropriate for the development of management technologies and maintenance of information. Research community every day focuses on the analysis of big data, as there is a great amount of information coming from platforms that analyze opinions about products, services, articles, experiences, travel. Not only digital, but from printed information from newspapers, accounting documents, invoices, stories, books, statistical data, surveys, countless places where you can generate any type of textual multi-information.

The accumulation of large volumes of textual information becomes an arduous task to try to generate an analysis manually, in addition to the need for research to get the computer interpret natural human language (NL). Thanks to this need and many others, Text Mining (TM) is created, whose purpose is to create a structuring in the information for future analysis, [1], defines it as the set of different techniques to find hidden patterns that allow generating knowledge. Thanks to the diversification of the use that generates, it is considered a multifunctional research tool based on: Information retrieval [2], mention that it is an activity that tries to obtain some relevant information resources for the need of information from large collections. Data mining (DM) [3], conclude that are the techniques that enable data exploration to get useful relation and correlation to retrieve valuable information from data. Machine learning [4], clarify that it is the computer technology that relies on algorithms to learn from training data in addition to other fields such as linguistic computing and statistics. It is important to note that TM is similar to DM with the difference that TM is in charge of structuring textual data for its processing.

II. THE TM PROCESS

TM is composed of a series of elements that allow its application and development. Figure 1 shows the TM process in detail.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10507

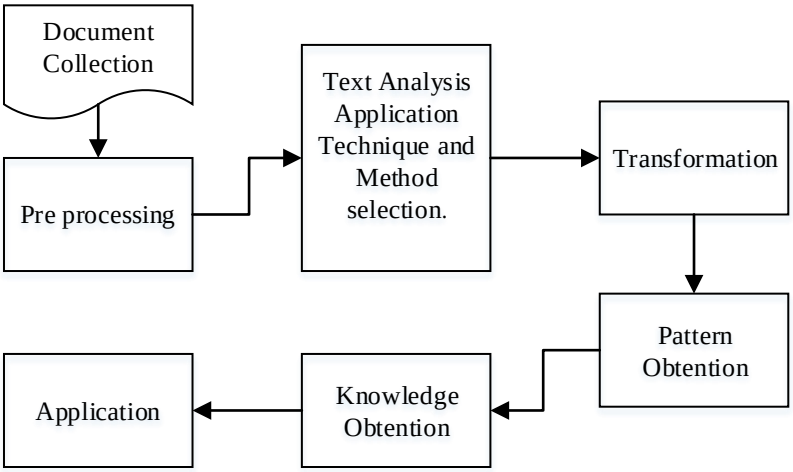

Fig 1. The TM process

Table 1 shows in detail the elements description of the overall TM process.

TABLE I DESCRIPTION OF THE OVERALL TM PROCESS

Name	**Description**
Document collection	The group of unstructured elements that will be analyzed may include web pages, comments, stories, books, pdfs, etc.
Pre-processing	It consists of formatting the information basically by filtering the information content, leaving useful elements structured for the process. It consists of 3 elements: A. Tokenization: [5], conclude that it is the process of dividing the textual flow into words, terms and symbols to meaningful elements for TM. B. Stemming: [6], manifest that is the process of derivation that returns words to their root word of origin containing its meaning. C. Stop Words: [7], described as the elimination of pre-established common words, with the aim of improving TM performance.
Text analysis selection of technique and method.	[8], define that there are two types of algorithms that allow the development of TM: A. Supervised learning algorithms: are those used when there is a set of predictors to predict a target variable. B. Unsupervised learning algorithms: are those that do not use an objective value to train their models, basically they use predictors to reveal hidden structures in the data.
Transformation	It is the process generated from the application of the technique.
Pattern obtention	It is the result of information from the application of the method.
Knowledge obtention	It is the final result of the application of TM.
Applications	The branch where the TM will be applied.

III. TM TECHNIQUES

There are different techniques where TM is used. Table 2 shows the common techniques.

TABLE 2 COMMON TM TECHNIQUES

Information Retrieval	[9], mention that it is the action of finding material of an unstructured nature (texts) that satisfies the need for information within large collections of data, stored in computers. Information that is searched every day in the search engines of the web can be seen in a practical way.
Information Extraction	[10], determine that it tries to extract predefined data types from a document, furthermore it aims to identify the most important objects by extracting the most important information. The idea is to establish relationships between words.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10507

Natural language processing (NLP)	[11], refers to a research area that explores how computers can be used to understand and manipulate text or speech in natural language to do useful things, basically to achieve an understanding between human language and computer language.
Clustering	[12], propose that it consists in discovering the underlying structure of a data set by partitioning the data into groups, similar elements with similar ones and non-similar ones with those that are different.
Text summarization	[13], state that they are those that automatically produce a summary that contains important sentences and include relevant information from the original document. There are two types of approaches A. Extractive Automatic Text Summarization (EATS): [14], suggest that their goal is to produce a brief text summary by extracting the most relevant sentences from a document, thus generating more fluent summaries than abstract approaches. B. Abstractive Automatic Text Summarization (AATS): [15], describes these as summaries that contain novel phrases not found in the source.

IV. TM APPLICATIONS

There are different applications for TM. Table 3 shows the TM applications in detail.

TABLE 3 TM APPLICATIONS

Sentiment analysis:	[16], argue that it is the way to know the user's opinion towards some kind of product or service besides being in direct contact with the end user. There are 3 types of emotions: positive, negative and neutral. It is commonly used by commercial companies, banks, universities, etc., in order to know what customers, think of their products or services, with the aim of improving them.
Classification of sentiments	[17], conclude that it is a speaker or writer attitude determination with respect to some specific topic, attitude can be any form of judgment or evaluation, of the author. It can also be used to predict sentiment opinion in texts, social media, books articles among other things. It can be applied by governments to generate political ideologies. it can also help in reviews of various subjects.
Argument text Mining	[18], mention that it is in charge of automatically detecting argumentative structures in textual discourse including essays, web forums, the structures can be generated automatically, it can generate arguments for and against.
Blog Mining	[19], concludes that it is the process of searching and analyzing blogs to generate additional information that would otherwise not be possible to find by examining only a single blog, through these analyses statements can be made, it tries to generate information from different sources.
Email Mining	[20], ratify that it is about facilitating the best use of emails and explore the commercial potentials of emails
Web mining	[21], state that it is a process activity that consists of extracting interesting patterns in user usage data records on the web. Elements that make up Web Mining: A. Resource search: It consists of retrieving expected web documents. B. Selection and pre-processing of information: It involves automatic selection and pre-processing of specific information from retrieved web resources. C. Generalization: It is the one that automatically discovers general patterns in web sites. D. Validation: Interpretation of the extracted patterns.
Social Media	[22], state through social media people can manifest different points of view, therefore with TM it has been possible to incorporate different forms of e-learning regardless of the fact that the opinions are not grammatically well written, thanks to TM. The adaptation of social media (Twitter, Facebook, etc.) has made it possible to generate tools that make possible to obtain the implicit knowledge within the information.
Medical Science	[23], conclude that are computational techniques and information technologies for managing large biomedical data repositories and discovering useful patterns, as well as gaining knowledge from them. In addition, they have made it possible to generate

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10507

	storage, retrieve, share, and manage multimedia. They have made it possible to describe various biological information related to drug discovery and patient care through statistical methods, and can be used to construct genetic pathways to provide a mapping into existing medical ontologies.

V. CONCLUSION

The application of Text Mining has no limit due to its great adaptability, it can be applied in almost any area of organizations, public or private, that handle textual information. Obtaining patterns of information extracted in an unstructured way allows the creation of new knowledge through the application of TM in different areas of daily life.

The choice of method in the application of TM will depend on the area where it is to be applied. TM is a powerful low-cost tool, it brings great benefits in the translation of texts into computer language, it favors the creation of new knowledge, it can be applied in the medical field as a tool for the calculation of medical treatments. When there are large amounts of information based on textual information, the application of TM allows this information to be evaluated to extract valuable information.

During the TM process there are great challenges to give and evaluate the information in order to understand it, being natural language processing the most important, understanding human language and transforming it into computational language is an arduous tool in which we continue to work every day. It is important to continue developing new algorithms that favor the application of TM. Thanks to TM, text summaries can be made in an automatic and easy way, they are tools that facilitate readers life in a significant way promoting the reduction of reading time by generating automatic summaries.

With tools such as sentiment analysis, companies have been able to understand more the taste and opinion of consumers. Companies like Amazon, eBay have managed to increase their sales massively in a short time because they focus on the needs of their customers.

Banks have managed to know the behavior of their employees and know opinions of their service with the application of surveys to their customers.

Medically it allows laboratories to run observations in large databases allowing the discovery of patterns that favor the creation of new medicines or even favoring drug trends, they have also managed to reduce costs in designing new medicines. In social media, it allows the analysis of user behavior and can also help to decipher what are the trends of the most prominent topics.

In web pages allows to achieve indexing in search engines thanks to keywords. On the other hand, it allows to search for important information in large quantities of web pages in an efficient manner. The idea of this research was to present a general overview of TM nowadays.

REFERENCES

[1]. Quillo-Espino, J. Romero-González, R. M., & Paulin Martinez, F. J. (2019). Text Mining Preprocessing In Times of Python vs MVCS.
[2]. Guo. J, Fan, Y., Pang, L. Yang, L., Ai, Q., Zamanic, H., Wu, C., Croftc, B., Cheng, X. (2019). A Deep Look into neural ranking models for information retrieval. https://doi.org/10.1016/j.ipm.2019.102067
[3]. Yu Sheng, S., &·Sheng Yi, W. (2020). Applying data mining techniques to explore user behaviors and watching video patterns in converged IT environments. https://doi.org/10.1007/s12652-020-02712-6
[4]. Zhai, X., Yin, Y. Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. 56(11), pp. 111-151. https://doi.org/10.1080/03057267.2020.1735757
[5]. Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenizacion tool an analysis. 3(1), pp.37-46. DOI:10.5121/acii.2016.3104 37
[6]. Maylawati, D. S., Zulfikar, W. B., Slamet, C., Ramdhani, M. A., & Gerhana, Y. A. (2018). An improved of stemming algotirhm for mining indonesian text with slang on social media. 6th International Conference on Cyber and IT Service Management (CITSM). doi: 10.1109/CITSM.2018.8674054.
[7]. Alshanik, F., Apon,A., Herzog, A., Safro, ll., & Sybrandt, J. (2020). Accelerating text mining using domain specific stop Wordlist. DOI:10.1109/BigData50022.2020.9378226
[8]. Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., Yeganegi, M. r. (2020). Text mining in Big data analytics. doi:10.3390/bdcc4010001
[9]. Poelmans, J., Ignatov, D. I., Viaene, S., Dedene, G., & Kuznetsov, S.O. (2012). Text Mining Scientific Papers: A Survey on FCA-Based Information Retrieval Research. Vol.7377, pp.273-287. https://doi.org/10.1007/978-3-642-31488-9_22
[10]. Salloum S, Mostafa A, Monem A, Shaalan K (2018) Using text mining techniques for extracting information from research articles. https://doi.org/10.1007/978-3-319-67056-0_18
[11]. Chowdhury, G. G. (2005). Natural Languaje processing. 37(1), pp.51-89. https://doi.org/10.1002/aris.1440370103
[12]. Vo-Van, T., Nguyen-Hai, A., Tat-Hong, M. V. & Nguyen-Trang, T. (2020). A new clustering algorithm and its application in assessing the quality of underground wáter. https://doi.org/10.1155/2020/6458576
[13]. Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E. Syukur, A., Affandy, A., & Setiadi, D. R. I. (2020). Review of automatic text summarization techniques & methods. https://doi.org/10.1016/j.jksuci.2020.05.006
[14]. Diao, Y., Lin, H., Yang, L., Fan, X., Chu, Y. Wu, D., Zhang, D., & Xu, K. (2020). CRHASum: extractive text summarization with contextualizedrepresentation hierarchical-attention summarization network. https://doi.org/10.1007/s00521-019-04638-3

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10507

[15]. Dong, Y., Wang, S., Gan, Z., Cheng, Y., Cheung, J. C. Y., & Liu, J. (2020). Multi-Fact Correction in Abstractive Text Summarization. arXiv:2010.02443

[16]. Vangara, R. V. B., Thirupathur, K., & Vangara, S. P. (2020). Opinion mining classification usign naive bayes algorithm. 9(5), pp.495-498. DOI: 10.35940/ijitee.E2402.039520

[17]. Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. https://doi.org/10.1016/j.dss.2009.09.003

[18]. Daxenberger, J., Schiller, B., Stahlhut, C., Kaiser, E., & Gurevych, I. (2020). ArgumenText: Argument Classification and Clustering in a Generalized Search Scenario. https://doi.org/10.1007/s13222-020-00347-7

[19]. O´Leary, D. E. (2011). Blog mining and extensión: from each according to his opinión. Decision Support Systems. 51(4), pp. 821-830. https://doi.org/10.1016/j.dss.2011.01.016.

[20]. Guanting, T, Jian, P., & Wo-Shun, L. (2015). Email Mining: Tasks, Common Techniques, and Tools. Obtenido el 5 de mayo de 2021 desde: https://www.cs.sfu.ca/~jpei/publications/EmailMining-KAIS.pdf.

[21]. Sivaramakrishnan, J., & Balakrishnan, V. (2009). Web Mining Functions in an Academic Search Application. 13(3), pp. 132- 139.

[22]. Salloum, S., Al-Emran, M., & Shaalan, K. (2017). A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering. 2(1), pp.127-133.

[23]. Chen, H. Fuller, S. S., Friedman, C. Hersh, W. (2005). Knowledge management, data mining, and text mining in medical informatics. https://doi.org/10.1007/0-387-25739-X_1